Business Analytics (110-1)

Assignment 3

Due: 9:00 am, Tue 23-Nov-2021

B08701119 工管三 張廷鋒

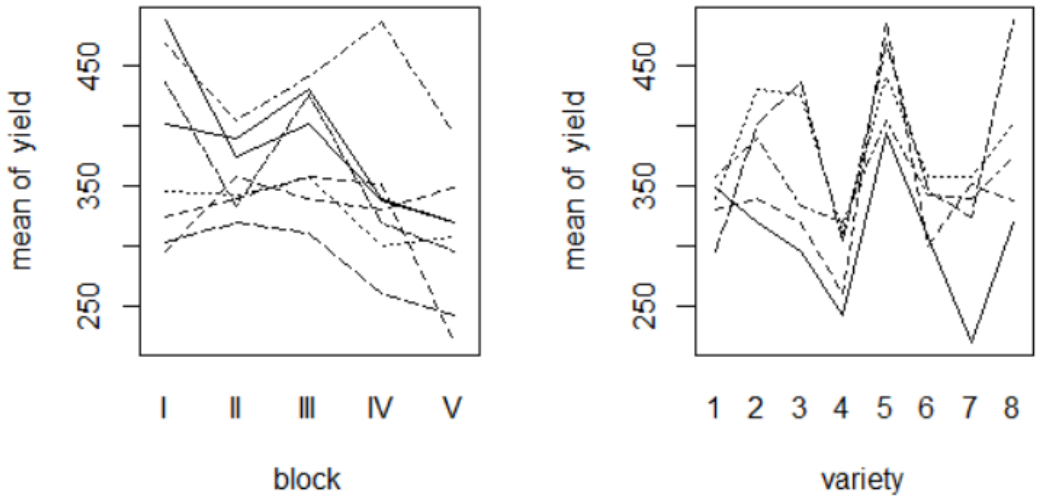
1.

A farmer wants to conduct an experiment to compare eight varieties of oats. The farmer knows that the growing area is heterogeneous so he decided to group the area into five blocks. He randomly plants each variety of oats in each block and records the yields accordingly. Dataset oatvar is the experiment result.

(a) What kind of experiment is this farmer using?

Two-way ANOVA

(b) Is there any interaction effect between the variety of oats and the growing area block?



查看interaction plot，發現兩張圖的直線都不平行，因此認為interaction effect存在

(c) Conduct a hypothesis test to determine whether the yield of oats is affected by different varieties at 5% significance level.

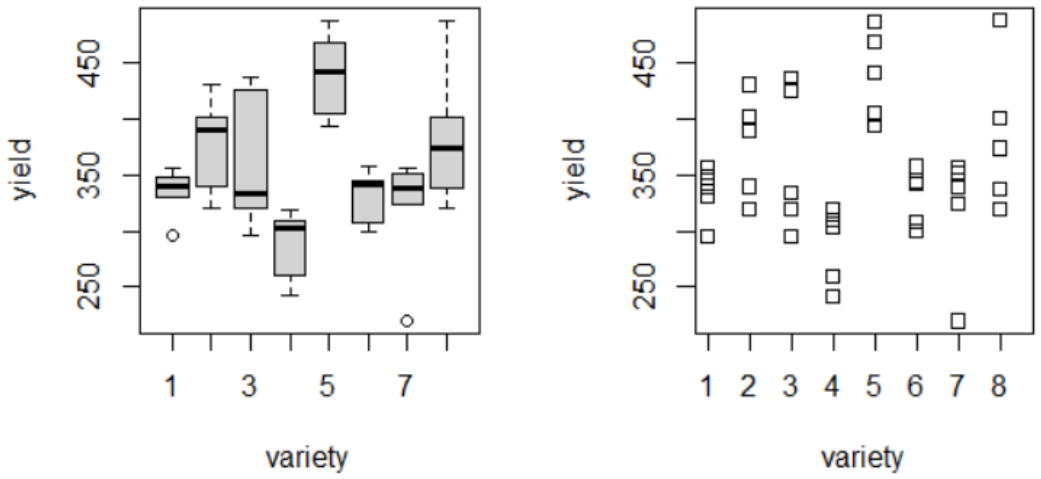
H0: means in all 8 levels are all the same

H1: at least one mean differ

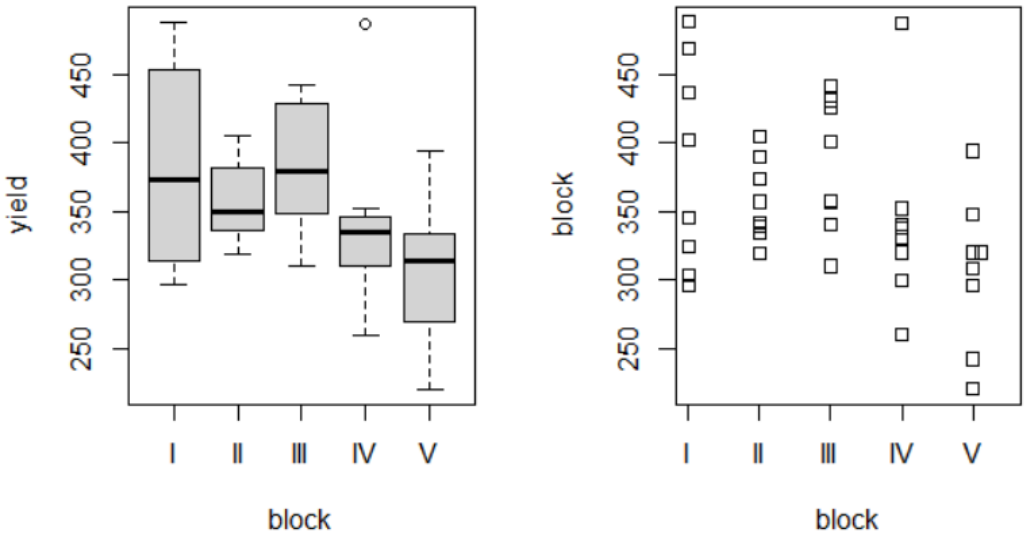
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         334.40      21.04  15.894  < 2e-16 ***
factor(variety)2     42.20      29.75   1.418  0.16578
factor(variety)3     28.20      29.75   0.948  0.35036
factor(variety)4    -47.60      29.75  -1.600  0.11949
factor(variety)5    105.00      29.75   3.529  0.00129 **
factor(variety)6     -3.80      29.75  -0.128  0.89918
factor(variety)7    -16.00      29.75  -0.538  0.59449
factor(variety)8     49.80      29.75   1.674  0.10394
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 47.05 on 32 degrees of freedom
Multiple R-squared:  0.5226,    Adjusted R-squared:  0.4181
F-statistic: 5.004 on 7 and 32 DF,  p-value: 0.0006568
```

觀察上圖執行one-way ANOVA的結果，可以看到ANOVA table的p-value為0.0006568，顯著小於significance level，因此可以推翻H0，認為different varieties會影響yield。

(d) Check the diagnostics. Is there any unusual findings?



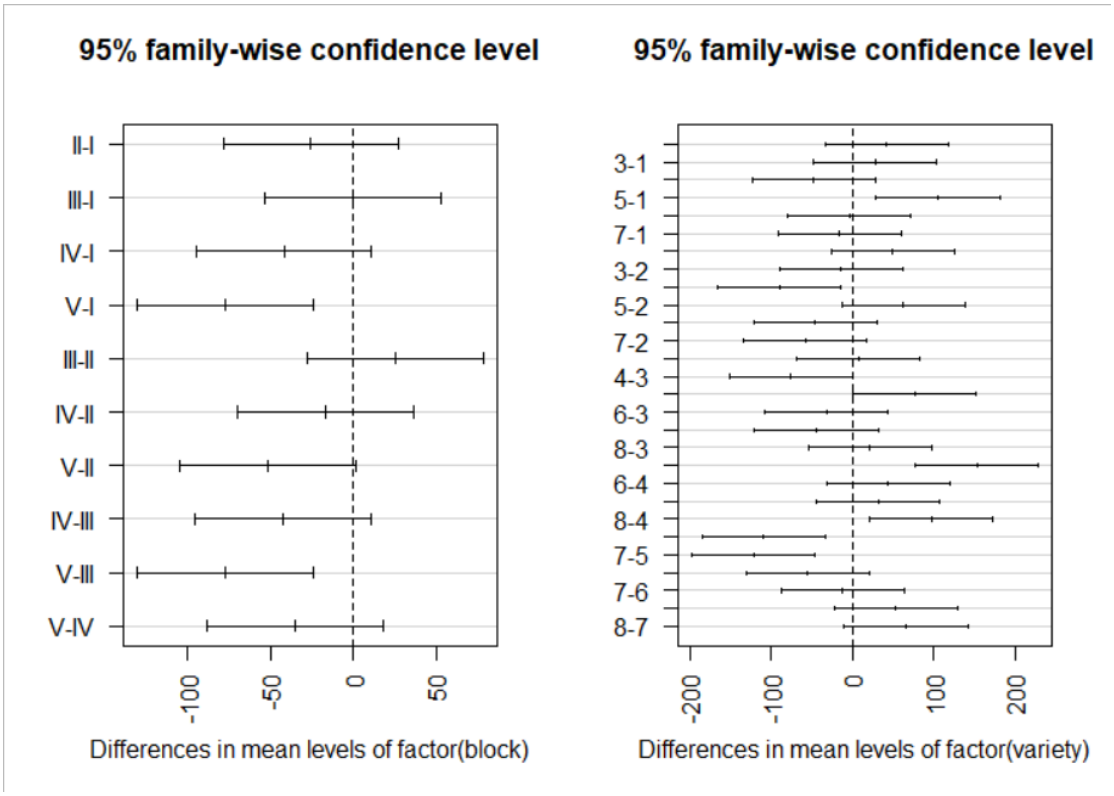
觀察上圖box plot，可發現不同variety之間yield的分布的確不同，與模型結果一致。但我們也可發現分布大約分為三種型態，第一種是variety為1、3、6、7；第二種為2、8，第三種與第四種分別為4和5。如果屬於同種類的variety，對yield的影響差異相對沒那麼大。



觀察block的boxplot可以發現，只有block5的分布與其他四個block明顯不同。

(e) Is it necessary to perform multiple comparisons? If yes, carry out the comparison following the structure and procedure presented in the lecture.

是，由於block與variety的影響都是顯著的，因此應該執行multiple comparison來確定兩兩之間的關係。



➔ block —
Tukey結果顯示除了 I-V、III-V, 其他pairs的差距並不顯著。
所以 I-V、III-V 是對yield造成影響的主要來源。
➔ variety —
Tukey結果顯示除了 1-5、2-4、3-7、4-8、5-6、5-7，其他pairs的差距並不顯著
所以 1-5、2-4、3-7、4-8、5-6、5-7 是對yield造成影響的主要來源。
2.
Detergent manufacturer frequently makes claims about the effectiveness of their products. A consumer-protection service decided to test the five best selling brands of detergent, where each manufacturer claims that its product produces the "whitest whites" in all water temperatures. The experiment was conducted in the following way. One hundred fifty white sheets were equally soiled. Thirty sheets were washed in each brand – 10 with cold water, 10 with warm water, and 10 with hot water. After washing, the "whiteness" scores for each sheet were measured with laser equipment.
Dataset detergent is the experimental result.

Column 1: water temperature code
Column 2: scores for detergent 1 (first 10 rows = cold water, middle 10 rows = warm, last 10 = hot)
Column 3: scores for detergent 2 (same format as column 2)
Column 4: scores for detergent 3 (same format as column 2)
Column 5: scores for detergent 4 (same format as column 2)
Column 6: scores for detergent 5 (same format as column 2)

Is there sufficient statistical evidence to infer that there are differences in whiteness scores between the five detergents, differences in whiteness scores between the three water temperatures, or interaction between detergent and temperatures?

```
Analysis of Variance Table

Response: Score
                     Df  Sum Sq Mean Sq F value    Pr(>F)
Detergent             4  2967.4  741.86  6.7170 5.830e-05 ***
Temperature           2  3937.1 1968.54 17.8238 1.351e-07 ***
Detergent:Temperature   8  2452.1  306.51  2.7752   0.00714 **
Residuals           135 14910.0  110.44
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test 1：differences in whiteness scores between the five detergents
H0: the means of the 5 levels of Detergent are equal
H1: At least two means differ

Test 2：differences in whiteness scores between the three water temperatures
H0: the means of the 3 levels of Temperature are equal
H1: At least two means differ
Test 3：interaction between detergent and temperatures
H0: Detergent and Temperature do not interact to affect the mean score
H1: Detergent and Temperature do interact to affect the mean score

觀察上圖two-way ANOVA執行結果，Detergent的p-value為5.830e-05 < 0.05，因此我們可以推翻Test 1的H0，認為differences in whiteness scores between the five detergents是存在的。Temperature的p-value為1.351e-07 < 0.05，因此我們可以推翻Test 2的H0，認為differences in whiteness scores between three water temperatures是存在的。Detergent : Temperature的p-value為0.00714 < 0.05，因此我們可以推翻Test 3的H0，認為interaction between detergent and temperatures是存在的。

3.
When car dealers lease a car, how do they decide what to charge? One answer, if you've got a lot of unpopular cars to move, is to charge whatever it takes to get the cars off the lot. A different answer considers the so-called "residual price" of the car at the end of the lease. The residual price of a leased car is the value of this car in the used-car market.

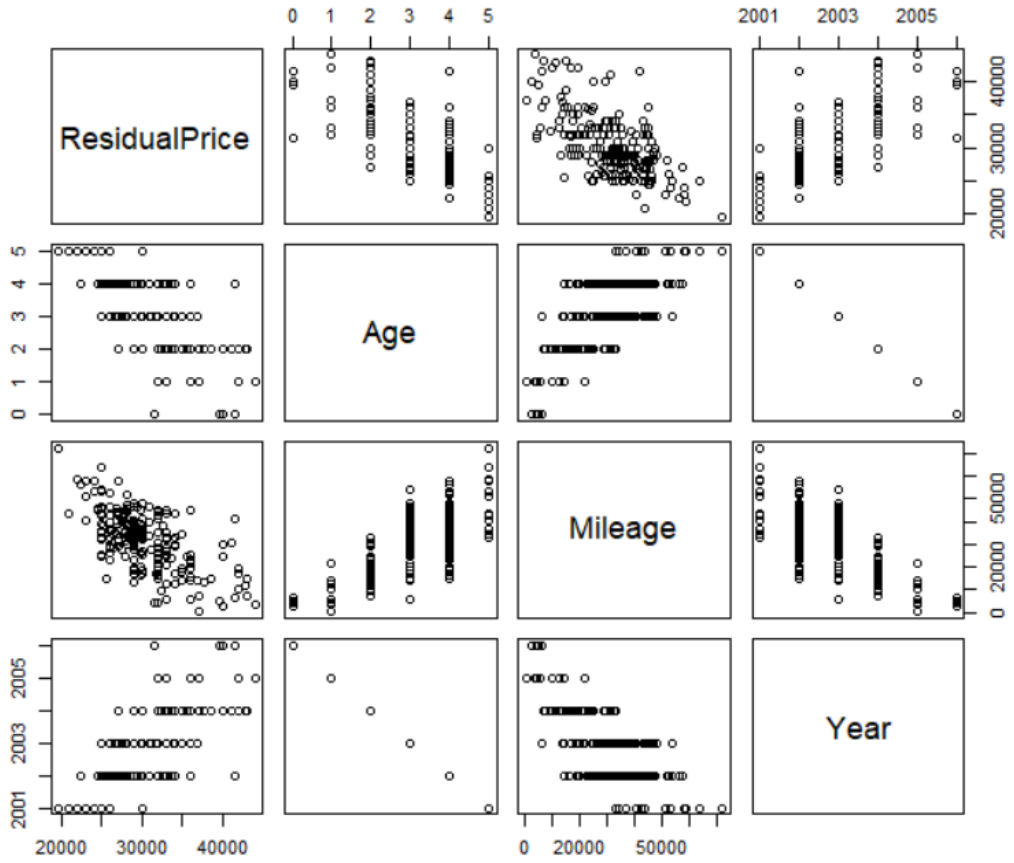
How should we estimate the residual price of a car? The residual price depends on how much the car was worth originally, such as the manufacturer's list price. Let's take this off the table by limiting our attention to a particular type of car. Let's also assume that we are looking at cars that have not been damaged in an accident.

What else matters? Certainly, the age of the car affects its residual price. Customers expect to pay less for older cars, on average. Older cars have smaller residual price. The term of the lease, say 2 or 3 years, has to cover the cost of the ageing of the car. Another factor that affects the residual price is the type of use. An older car that is in the great condition might be worth more than a newer car that has been heavily driven. It seems as though the cost of a lease ought to take both duration and use into account.

Dataset used_bmw lists 218 BMW's popular 3-series.

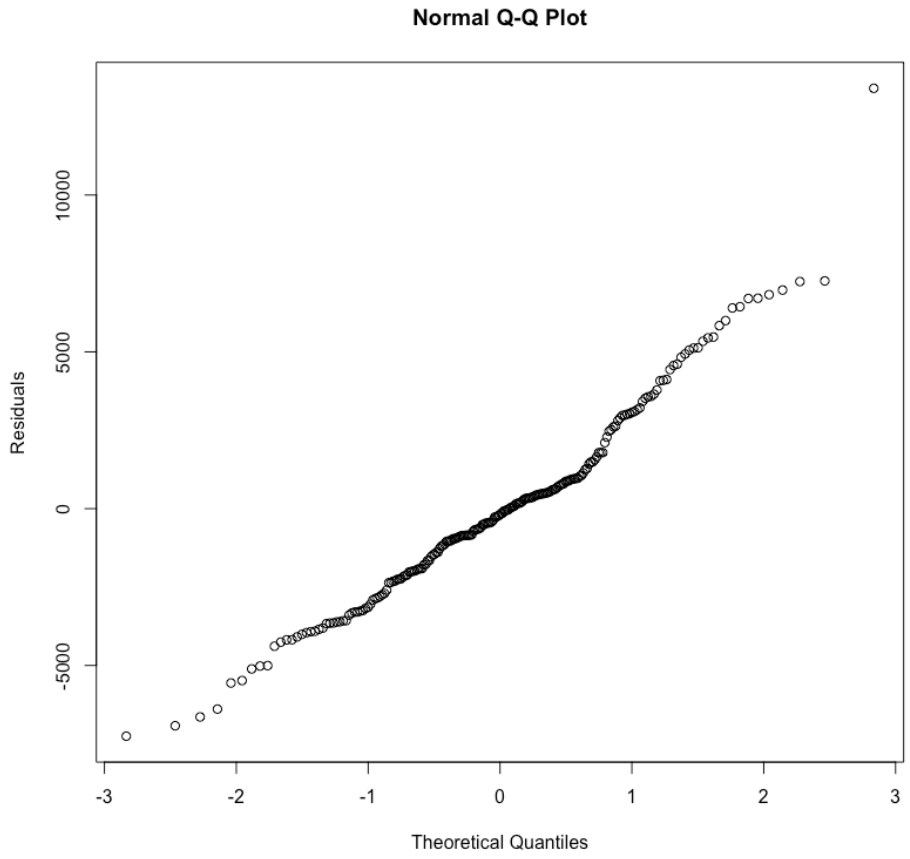
(a) Check scatterplots of the variables. Do the relationships appear straight enough

to permit using multiple linear regression with these variables?

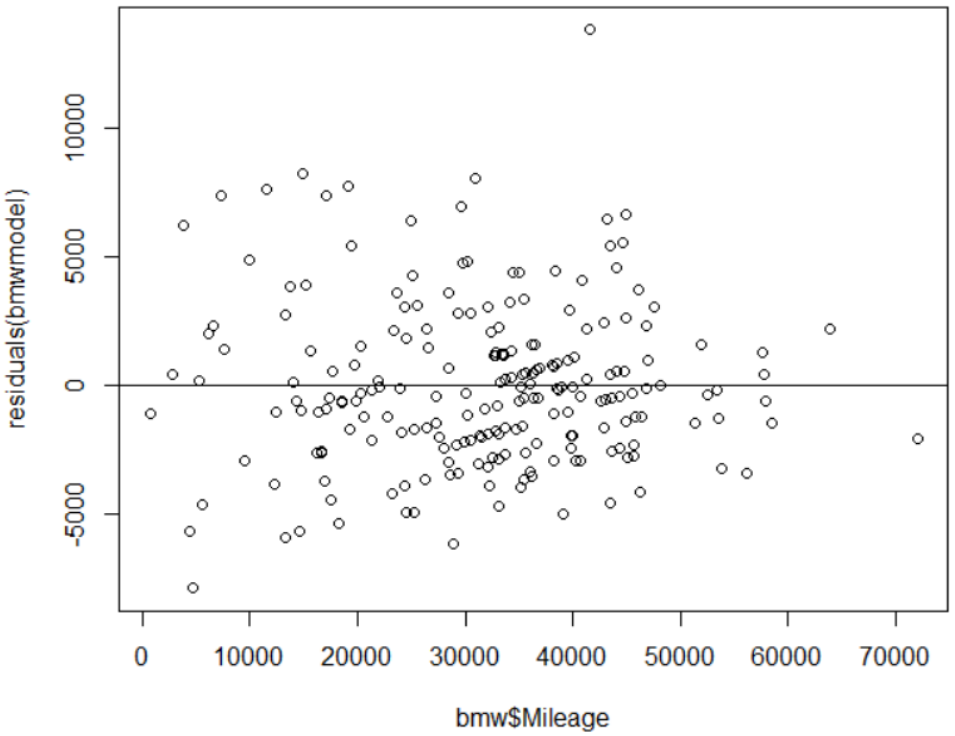


觀察上圖，可看到Residual Price和三個變數都有相關性，可以使用multiple linear regression model。其中，Age是從Year推算出來，因此建模時為了避免多重共線性，只需要使用其中一個變數age就好了。

(b) Fit the appropriate multiple linear regression model. Does this model meet the assumptions for multiple linear models?


Normal Q-Q Plot
Residuals
10000
5000
0
-5000
-3
-2
-1
0
1
2
3
Theoretical Quantiles

(c)


residuals(bmwmodel)
10000
5000
0
-5000
0
10000
20000
30000
40000
50000
60000
70000
bmw$Mileage

QQ plot的分布是常態的，但從Residual Plot的分布來看，並不符合假設。

(d) Build confidence intervals for the partial effects of age and mileage.

```
> confint(bmwmodel)
                    2.5 %        97.5 %
(Intercept)  3.753067e+04  4.221872e+04
Mileage     -2.055417e-01  2.263640e-03
Age         -2.536180e+03 -8.823816e+02
Mileage:Age -3.332859e-02  2.036568e-02
```

從此圖可看出，mileage斜率信賴區間下限為-0.205，上限為0.00226。Age斜率信賴區間下限為-2536.1，上限為-882.3。

(e) Summarize the results of your model. Recommend terms for leases that cover the costs of ageing and mileage.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.987e+04  1.189e+03  33.531  < 2e-16 ***
Mileage     -1.016e-01  5.271e-02  -1.928   0.0552 .
Age         -1.709e+03  4.195e+02  -4.074  6.5e-05 ***
Mileage:Age -6.481e-03  1.362e-02  -0.476   0.6347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3185 on 214 degrees of freedom
Multiple R-squared:  0.5109,    Adjusted R-squared:  0.504
F-statistic: 74.51 on 3 and 214 DF,  p-value: < 2.2e-16
```

Global usefulness test

H0 : All coefficients equal 0

H1: At least one coefficient != 0

先觀察Global usefulness test的部分，p-value為2.2e-16，代表可以推翻全部係數為0的虛無假設，證明模型是有效的。

接著觀察t-test的部分，Intercept與Age的斜率皆為顯著，Mileage的斜率p-value為0.0552，只比0.05大一些，仍在可接受範圍內。此外，interaction項的p-value為0.6347，顯示interaction的影響為不顯著。

Terms: ResidualPrice = 3987 -0.1016 * Mileage -1709 * Age

(f) Do you have any caveats that should be pointed out with your recommended terms? For example, are there any evident lurking variables?

```
> vif(fit)
    Mileage         Age Mileage:Age
   9.550654    4.088656   16.376625
```

By verifying VIF, we can conclude that there is no multicollinearity in this model

藉由VIF驗證，我們可以認定模型無多重共線性。